

# 自适应拜占庭鲁棒的差分隐私联邦学习

王玉画<sup>1</sup>, 张沁楠<sup>1\*</sup>, 邱望洁<sup>1</sup>, 柴子川<sup>2</sup>, 高胜<sup>3</sup>, 朱建明<sup>3</sup>, 童咏昕<sup>4</sup>, 郑志明<sup>1</sup>

1. 北京航空航天大学人工智能学院未来区块链与隐私计算高精尖创新中心,北京 100191

2. 北京航空航天大学机械工程及自动化学院,北京 100191

3. 中央财经大学信息学院,北京 100081

4. 北京航空航天大学计算机学院,北京 100191

\* 通信作者. E-mail: zhangqn@buaa.edu.cn

收稿日期: 2025-05-29; 修回日期: 2025-07-25; 接受日期: 2025-08-21; 网络出版日期: 2025-10-16

国家重点研发计划(批准号: 2022ZD0116800)、国家自然科学基金(批准号: 62372493, 62141605)、北京市自然科学基金(批准号: Z230001)、中国博士后科学基金(批准号: 373500)和北航敢为行动计划(批准号: KG16336101)资助项目

**摘要** 联邦学习能够在数据本地私有的前提下实现跨设备协同训练,但在实际应用中仍面临隐私泄露和投毒攻击两大安全瓶颈。差分隐私和鲁棒聚合作为各自领域的主流防御方法,在耦合应用时存在固有的内在冲突:差分隐私所注入的随机噪声一方面放大了良性更新的分布方差,另一方面也为恶意更新的系统性偏移提供了掩护,使得两者难以有效区分。为此,本文提出自适应拜占庭鲁棒差分隐私联邦学习方法 AByzDPFL,旨在通过降低噪声维度并放大模型间的几何差异来提高可区分度。在客户端侧,基于 Fisher 信息的私有选择机制动态筛选关键参数坐标,仅在该低维子空间中注入噪声并完成更新,从而显著减少噪声维度,降低良性模型的方差。在服务器侧,使用谱嵌入降维突出本征几何结构,随后根据噪声尺度自适应聚类半径,从而包容受噪声扰动的良性模型,并有效剔除超出噪声范围的系统性偏移。进一步地,通过自适应中值范数裁剪抑制簇内的高幅值异常更新。理论分析证明了所提方法的隐私损失和收敛性上界,实验表明 AByzDPFL 在兼顾隐私性和鲁棒性的同时,性能优于现有的主流基线。

**关键词** 联邦学习, 差分隐私, 拜占庭鲁棒, 选择性更新, 噪声自适应

## 1 引言

联邦学习(federated learning, FL)<sup>[1]</sup>作为一种新兴的分布式机器学习范式,通过将模型训练任务下发到多个客户端,仅需上传本地模型参数或梯度而非原始数据即可实现全局模型聚合,在保障数据本地化的同时完成分布式联合建模与全局优化。随着数据隐私保护需求日益增长和数据孤岛问题日益凸显,联邦学习已在移动智能、医疗健康、金融风控等领域展现出显著优势,成为跨机构数据协同与隐私安全计算的关键技术<sup>[2~5]</sup>。

**引用格式:** 王玉画, 张沁楠, 邱望洁, 等. 自适应拜占庭鲁棒的差分隐私联邦学习. 中国科学: 信息科学, 2025, doi: 10.1360/SSI-2025-0232  
Wang Y H, Zhang Q N, Qiu W J, et al. Adaptive Byzantine-robust differentially private federated learning. Sci Sin Inform, 2025, doi: 10.1360/SSI-2025-0232

然而,在公开网络环境的实际部署中,隐私泄露风险与系统鲁棒性缺陷仍是阻碍联邦学习规模化应用的核心瓶颈。具体而言,中心服务器可能对客户端上传的模型更新实施逆向工程或推理攻击,从而反推出用户的敏感信息<sup>[6,7]</sup>。同时,恶意客户端可以通过上传篡改后的参数破坏全局模型收敛性,甚至植入隐蔽的后门攻击<sup>[8,9]</sup>。为进一步满足轻量化部署需求,差分隐私和鲁棒聚合分别成为了应对这两类问题的主流解决方法。其中,差分隐私(differential privacy, DP)<sup>[10]</sup>保护通过本地梯度中注入随机噪声以模糊敏感信息,但其默认假设所有客户端均为诚实参与者,无法抵御恶意节点通过噪声掩盖投毒攻击的行为。而拜占庭鲁棒(Byzantine robustness, BR)<sup>[11]</sup>聚合虽能剔除异常参数的影响,却需要服务器直接在明文梯度上进行操作,存在联邦学习隐私悖论。因此,在兼顾隐私保护与投毒防御的场景中,一种自然的想法是将现有的两类技术直接结合,但这种简单的耦合存在如下冲突。(1) 噪声扩散。在全参数空间均匀注入噪声会使良性更新的总体方差随模型维度 $d$ 线性增长,导致基于统计距离的鲁棒检测频繁误判。(2) 隐蔽攻击。攻击者可利用噪声的统计特性,在噪声量级及方向相似的范围内施加系统性偏移,使得固定阈值的聚类投毒检测难以可靠区分。

针对上述问题,本文提出了自适应拜占庭鲁棒的差分隐私联邦学习方法(adaptive Byzantine-robust differentially private federated learning, AByzDPFL),其核心在于通过压缩噪声作用于空间并放大本地模型间的几何结构差异,从而提升对随机噪声漂移与恶意系统性偏移的区分能力。在客户端侧,基于经验Fisher信息量的私有选择机制,筛选出前 $k$ 个关键参数坐标,仅在该 $k$ 维子空间注入高斯噪声并完成更新,使有效噪声维度由 $d$ 降至 $k$ ,在严格满足DP的同时显著降低了良性更新方差。在服务器侧,首先对高维带噪更新进行谱嵌入降维,保留模型间的本征几何结构并削弱噪声的扩散性。其次,根据噪声尺度为每个模型自适应聚类半径,确保在聚类过程中既能包容正常的噪声漂移,又能有效区分出超出噪声范围的系统性偏移,从而有效提取出最大良性簇。此外,对簇内模型施加自适应中值范数裁剪,抑制残留的高幅值异常。理论分析表明,AByzDPFL在 Rényi DP 约束下具备严格的隐私损失上界,并在非凸优化场景下获得可量化的收敛性分析。丰富的实验评估表明,AByzDPFL不仅能抵御成员推断攻击,还在4种典型的投毒攻击中超越主流基线方法的性能。本文的主要贡献总结如下。

- (1) 提出基于参数信息量的选择性扰动更新策略。仅对高信息量参数注入差分隐私噪声,剩余低信息坐标本次迭代冻结,从而降低良性模型的方差以提高可用性。
- (2) 提出多级自适应的噪声感知鲁棒聚合规则。结合谱嵌入降维、噪声自适应聚类与中值范数裁剪,实现对带噪更新的高效异常检测与剔除,全面提升对多种投毒攻击的防御能力。
- (3) 提供隐私性和收敛性两方面严格的理论保证。首先证明算法在多轮迭代下隐私组合边界,并推导全局收敛性保证,定量刻画稀疏扰动与鲁棒聚合下隐私与性能的协同关系。
- (4) 通过各种对比和消融实验验证方法的有效性,并表明其在不同恶意客户端比例和非独立同分布(Non-IID)场景下具备良好的适应性。

本文其余部分的结构组织如下。第2节首先介绍了相关工作。第3节对本文的理论知识进行了介绍。第4节介绍了系统模型、威胁模型和设计目标。第5节介绍了自适应拜占庭鲁棒的差分隐私联邦学习方法的详细设计。第6节对隐私性和收敛性进行了严格的理论证明。第7节提供了对本文所提方法的实验验证。第8节对全文进行了总结。

## 2 相关工作

**差分隐私联邦学习。**为保护联邦学习中的用户敏感信息,DP因其计算高效、通信开销低以及后处理不变性等优势,相较于同态加密<sup>[12]</sup>和安全多方计算<sup>[13]</sup>等加密方案,更适合于客户端侧的轻量化部署。Geyer等<sup>[14]</sup>首次在客户端级别引入高斯噪声,McMahan等<sup>[15]</sup>提出DP-FedAvg结合采样与矩会计<sup>[16]</sup>进一步优化了隐私预算,奠定了该方向的基础。随后,研究者主要聚焦于降低噪声对模型性能的影响。Wei等<sup>[17]</sup>设计分阶段的差分隐私联邦学习NbAFL,根据训练阶段自适应调节当前轮数全局梯

度所需的噪声强度. Andrew 等<sup>[18]</sup> 提出基于参数范数分布的自适应裁剪方法, 突破了固定裁剪阈值的性能瓶颈. Cheng 等<sup>[19]</sup> 通过模型规范化与本地更新稀疏化降低噪声维度, 验证了本地参数压缩对隐私预算的有效控制. Xu 等<sup>[20]</sup> 针对分类任务的 softmax 层参数特性提出本地保留敏感层策略, 避免标签规模扩大导致的噪声累积. 最近, Wang 等<sup>[21]</sup> 的 FedLAP-DP 进一步通过共享合成样本近似本地损失面, 显著提升了紧隐私预算下的收敛速度和性能.

**联邦学习投毒防御.** 联邦学习的开放参与特性使其易受恶意客户端的投毒攻击<sup>[22]</sup>. 早期防御工作多依赖统计距离, 如 Blanchard 等<sup>[11]</sup> 提出 Krum 通过最小化欧氏距离筛选更新, Yin 等<sup>[23]</sup> 的 TrimMean 对各维度截断极值以提升鲁棒性, 但在 Non-IID 场景下常失效. Fung 等<sup>[24]</sup> 提出 FoolsGold, 通过计算各模型更新之间的余弦相似度, 动态调整对全局模型的聚合权重. 随着攻击复杂化, Li 等<sup>[25]</sup> 提出的 LoMar 通过分析更新分布的动态阈值实现恶意行为识别, 而 Nguyen 等<sup>[26]</sup> 提出的 FLAME 则采用 HDBSCAN 聚类算法增强对异常更新的检测能力, 并引入高斯噪声干扰后门攻击. Bao 等<sup>[27]</sup> 则通过辅助数据子空间投影来分离诚实梯度与异常更新, 有效解决了 Non-IID 场景下的检测偏差问题. 此外, 一些方法聚焦于信任评估体系的构建. Cao 等<sup>[28]</sup> 提出 FLTrust, 利用本地模型在辅助数据集上生成参考梯度, 进而构建信任评分作为每个客户端的聚合权重. Chu 等<sup>[29]</sup> 提出融合主观逻辑与残差检测的声誉模型, 通过历史表现动态调整客户端声誉并残差分析可疑更新.

**隐私与鲁棒的协同防御.** 考虑到现实场景的安全需求, 近年来出现了许多面向隐私与鲁棒协同的方案. Zhou 等<sup>[30]</sup> 结合权重截断与轻量级异常检测, 于两端动态注入高斯噪声, 实现资源受限场景下的隐私与安全并行. Yang 等<sup>[31]</sup> 在 PR-PFL 中融合 DP-SGD 与多任务个性化训练, 以鲁棒聚合和本地微调应对多种可用性攻击. 然而, 上述这些方法直接与服务器端的投毒检测策略结合, 并未考虑到噪声对鲁棒聚合产生的干扰, 使得对有毒模型剔除能力有限. Lan 等<sup>[32]</sup> 的 PRoBit+ 采用一位量化上传并在比特级别实现  $(\epsilon, 0)$ -DP 与拜占庭鲁棒, 通信开销极低但以一定精度损失为代价. 为了隔离噪声的干扰, Tang 等<sup>[33]</sup> 提出 PILE 框架, 利用可验证扰动和零知识证明同时实现模型保护与梯度正确性验证. Rathee 等<sup>[34]</sup> 的 ELSA 协议以相关性随机数替代传统交互, 提升恶意场景下的安全性与效率. Zhang 等<sup>[35]</sup> 引入可验证秘密共享与多级服务器架构, 在剔除异常更新后对秘密共享模型加噪, 取得了高攻击比例下的微小精度损失, 但其对多个次级服务器的部署成本过高.

总的来看, 上述协同方案要么未能充分考虑噪声对鲁棒聚合的内在冲突, 要么依赖复杂的加密与多方安全协议, 导致系统复杂度和通信开销显著上升. 如何在有限资源和实际部署场景下, 有机融合轻量级差分隐私保护与噪声自适应鲁棒聚合, 构建高安全性和高可用性的联邦学习方法, 仍是该领域亟须突破的核心科学难题.

### 3 理论基础

#### 3.1 差分隐私

差分隐私 (DP)<sup>[10]</sup> 旨在确保攻击者无法通过查询结果推断某条记录是否在数据集中. 具体而言, DP 保证查询结果对任意单一记录的变化不敏感.

**定义1**  $((\epsilon, \delta)\text{-DP}$ <sup>[10]</sup>) 随机化机制  $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ , 若对任意相邻数据集  $\mathcal{D}, \mathcal{D}' \in \mathcal{D}$  (它们最多仅在一个元素上不同) 以及任意可测输出子集  $O \subseteq \mathcal{R}$ , 满足  $\Pr[\mathcal{M}(\mathcal{D}) \in O] \leq e^\epsilon \Pr[\mathcal{M}(\mathcal{D}') \in O] + \delta$ . 其中,  $\epsilon > 0$  控制隐私损失, 决定输出的可区分程度;  $\delta$  则允许有极小概率的失败.

为了满足  $(\epsilon, \delta)\text{-DP}$ , 可以使用高斯机制, 即向查询结果中添加按  $\ell_2$  敏感度校准的高斯噪声.

**定义2** (敏感度<sup>[36]</sup>) 给定查询函数  $f$ , 其  $\ell_p$  敏感度定义为  $\Delta_f = \max_{\mathcal{D}, \mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_p$ , 其中  $\mathcal{D}, \mathcal{D}' \in \mathcal{D}$  是最多相差一个元素的相邻数据集,  $\|\cdot\|_p$  表示  $\ell_p$  范数, 通常选取  $\ell_1$  或  $\ell_2$ .

### 3.2 Rényi 差分隐私

Rényi 差分隐私 (Rényi differential privacy, RDP) [37] 基于 Rényi 散度  $D_\alpha(P\|Q)$  来度量输出分布的差异, 通过调整  $\alpha$  值控制隐私损失.

**定义3** (RDP [37]) 设随机机制  $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ , 对任意一对仅差一个样本的相邻数据集  $\mathcal{D}, \mathcal{D}'$  满足  $D_\alpha(\mathcal{M}(\mathcal{D})\|\mathcal{M}(\mathcal{D}')) = \frac{1}{\alpha-1} \log \mathbb{E}_{\theta \sim \mathcal{M}(\mathcal{D}')} \left[ \left( \frac{\mathcal{M}(\mathcal{D})(\theta)}{\mathcal{M}(\mathcal{D}')(\theta)} \right)^\alpha \right] \leq \epsilon(\alpha)$ , 则称  $\mathcal{M}$  满足  $(\alpha, \epsilon)$ -RDP.

下述引理给出了将  $(\alpha, \epsilon(\alpha))$ -RDP 转换为  $(\epsilon, \delta)$ -DP 的标准形式.

**引理1** ( $(\alpha, \epsilon(\alpha))$ -RDP 到  $(\epsilon, \delta)$ -DP 的转换 [37]) 若机制  $\mathcal{M}$  保证  $(\alpha, \epsilon(\alpha))$ -RDP, 则对任意  $\delta \in (0, 1)$ ,  $\mathcal{M}$  也保证  $(\epsilon(\alpha) + \frac{\log(1/\delta)}{\alpha-1}, \delta)$ -DP.

### 3.3 One-Shot Laplace

One-Shot Laplace 机制是一种一次性差分私有 Top- $k$  选择方法, 通过一次性 Laplace 噪声注入, 即可在单次排序中实现对 Top- $k$  选择的严格隐私保护, 极大地简化了多轮剥离的开销.

**定义4** 给定待选择的函数集合为  $\{f_1, \dots, f_d\}$ , 每个  $f_i$  的  $\ell_1$  敏感度为  $\Delta_f$ , 则 One-Shot Laplace 机制对每个  $i = 1, \dots, d$  加入独立噪声  $b_i \sim \text{Lap}(\lambda)$ :  $y_i = f_i(\mathcal{D}) + b_i$ , 再将  $\{y_i\}$  从小到大排序, 返回前  $k$  个索引及其对应的  $f_i(\mathcal{D})$  的近似值.

该机制在纯差分隐私和近似差分隐私两种情形下均有严格的隐私保证.

**定理1** (纯 DP 保证) 如果将拉普拉斯噪声尺度  $\lambda$  设置为  $\lambda \geq \frac{2k\Delta_f}{\epsilon}$ , 则 One-Shot Laplace 机制满足  $(\epsilon, 0)$ -差分隐私.

**定理2** (近似 DP 保证) 设  $\epsilon_2 \leq 0.2$ ,  $\delta_2 \leq 0.05$  且  $d \geq 2$ . 如果将 One-Shot Laplace 机制的噪声尺度  $\lambda$  设置为  $\lambda \geq \frac{8\Delta_f\sqrt{k \log(d/\delta)}}{\epsilon}$ , 则该机制满足  $(\epsilon, \delta)$ -差分隐私.

## 4 问题描述

### 4.1 系统模型

本文关注横向 (水平) 联邦学习场景中, 考虑一个由中心服务器和  $N$  个客户端组成的分布式系统. 服务器负责全局模型的初始化、参数下发以及聚合更新, 其并不直接访问客户端的原始数据, 而是基于各客户端上传参数执行鲁棒聚合算法. 每个客户端  $i$  持有本地数据集  $D_i = (x_{i,j}, y_{i,j})_{j=1}^{n_i}$ , 并参与模型训练以实现知识共享, 其中  $n_i$  表示第  $i$  个客户端数据样本的数量. 所有客户端的数据样本均位于相同的特征空间  $\mathcal{X}$  与标签空间  $\mathcal{Y}$ , 但由于数据收集来源和分布差异, 可能呈现 Non-IID 特征. 在第  $t$  ( $1 \leq t \leq T$ ) 轮通信中, 服务器依据采样率  $q$  随机选择客户端子集  $\mathcal{A}_t \subseteq \mathcal{C}$ , 其规模为  $m = \lceil qN \rceil$ . 服务器将当前全局模型参数  $w^t$  下发至所有选中客户端, 每个客户端利用本地数据  $D_i$  进行  $E$  轮局部训练后, 上传模型更新  $w_i^{t+1}$ , 并由服务器通过鲁棒聚合算法进行全局更新. 本文涉及的重要符号如表 1 所示.

### 4.2 威胁模型

假设客户端与服务器之间的通信由如 SSL/TLS 等加密协议保护, 且不存在外部窃听攻击者, 双方也不进行合谋攻击. 本文考虑一个更为强大且实用的威胁模型, 即非完全可信模型, 其中半诚实服务器和恶意客户端共同存在. 具体而言, 服务器是诚实但好奇的, 它遵循聚合协议, 但可能对客户端的隐私数据产生兴趣. 在接收本地模型更新后, 服务器可能尝试进行隐私推断攻击, 从而反向推断出客户端的敏感信息. 每个本地客户端可能是恶意的或已被控制的, 并可能发起非定向投毒攻击, 旨在破坏全局模型的训练, 导致在测试时全局模型产生较高的错误率. 为了简化描述, 本文将所有恶意客户端 (无

表 1 符号总结.

Table 1 Summary of symbols.

Symbol	Description
$\mathcal{C}, \mathcal{A}, \mathcal{S}$	Total clients, sampled clients, central server
$N, m, i$	Total number of clients, number of sampled clients, client index
$T, t; E, B$	Number of communication rounds, index of round; local epochs, local batch size
$d, d_l, k_0$	Total parameter dimension, parameters in layer $l$ , embedding dimension for spectral clustering
$p, k_l, k$	Sparsification ratio, selected coordinates in layer $l$ , total selected coordinates
$w^t, w_i^t, \tilde{w}$	Global model at round $t$ , client $i$ 's model at start of $t$ , locally perturbed model
$\mathcal{D}_i,  \mathcal{D}_i $	Dataset of client $i$ , the size of $\mathcal{D}_i$
$(\epsilon_1, \delta_1), (\epsilon_2, \delta_2)$	Privacy parameters for coordinate selection, privacy parameters for parameter update
$\lambda, \sigma$	Laplace noise scale for Top- $k$ , Gaussian noise standard deviation for model update
$F_{(j)}, \hat{F}_{(j)}, \tilde{F}_{(j)}$	Empirical Fisher score, normalized Fisher score, noisy Fisher score
$C, \Delta_F$	Gradient clipping threshold, $\ell_1$ -sensitivity of Fisher scores
$\gamma, W, D, U$	Kernel bandwidth, similarity matrix, degree matrix, spectral embedding matrix
$\text{MinPts}, \text{Eps}_i, \mathcal{Z}$	Minimum number of neighbors, neighborhood radius, benign candidate index set

论是原生恶意还是被控制的) 统称为恶意客户端. 此外, 假设恶意客户端的比例不超过总客户端数的 50%, 否则恶意客户端可能轻松操纵全局模型.

### 4.3 设计目标

本文所提方法旨在实现联邦学习模型隐私保护的同时, 有效抵抗拜占庭攻击. 设计目标如下. (1) 数据隐私性. 确保客户端本地模型更新在传输和聚合过程中保持其机密, 抵御诚实但好奇的服务器对参数的反向推断攻击. (2) 模型鲁棒性. 面对恶意客户端发起的非定向投毒, 能够及时检测并剔除异常更新, 维护全局模型训练过程的安全性和稳定性. (3) 模型可用性. 在引入隐私保护与鲁棒防御机制后, 保证全局模型正常收敛, 且精度损失维持在可接受范围, 与无防护场景下的性能尽可能接近. (4) 资源高效性. 考虑到跨设备场景中客户端资源受限, 本方法避免引入显著的计算或通信开销.

## 5 本文工作

针对联邦学习中的数据隐私性和模型鲁棒性的内在冲突, 本文提出一种自适应拜占庭鲁棒的差分隐私联邦学习方法, 整体框架如图 1 所示. 该方法主要包含选择性扰动更新 (selective perturbation update, SPU) 和噪声感知鲁棒聚合 (noise-aware robust aggregation, NARA) 两大关键模块, 下文分别对二者展开详细介绍.

### 5.1 选择性扰动更新 (SPU)

传统的全局梯度加噪方法会在全参数空间中均匀注入噪声, 这些随机噪声在所有参数维度间扩散, 导致方差上界随参数维度  $d$  线性增长 [38, 39]. 这种高维噪声不仅严重削弱了模型的收敛性能, 也为恶意的系统性偏移提供了一定的掩护. 一种自然的想法是直接仅在梯度幅值较大的坐标上注入噪声以压缩噪声维度, 但却容易泄露模型对特定输入或特征的敏感性, 攻击者可据此反推出本地数据的敏感统计特征, 威胁用户隐私. 为解决上述矛盾, 本文提出选择性扰动更新策略: 客户端首先基于经验 Fisher 信息量在本地私有地筛选出  $k$  个关键信息坐标, 然后仅对这  $k$  维参数注入高斯噪声并完成更新, 其余  $d - k$  维在本次迭代保持不变. 该策略在确保隐私保护的同时, 通过降低有效噪声方差, 最大限度地保留良性模型的主要几何结构, 从而保障服务器端对投毒攻击的检测能力和全局聚合精度. 整体流程如

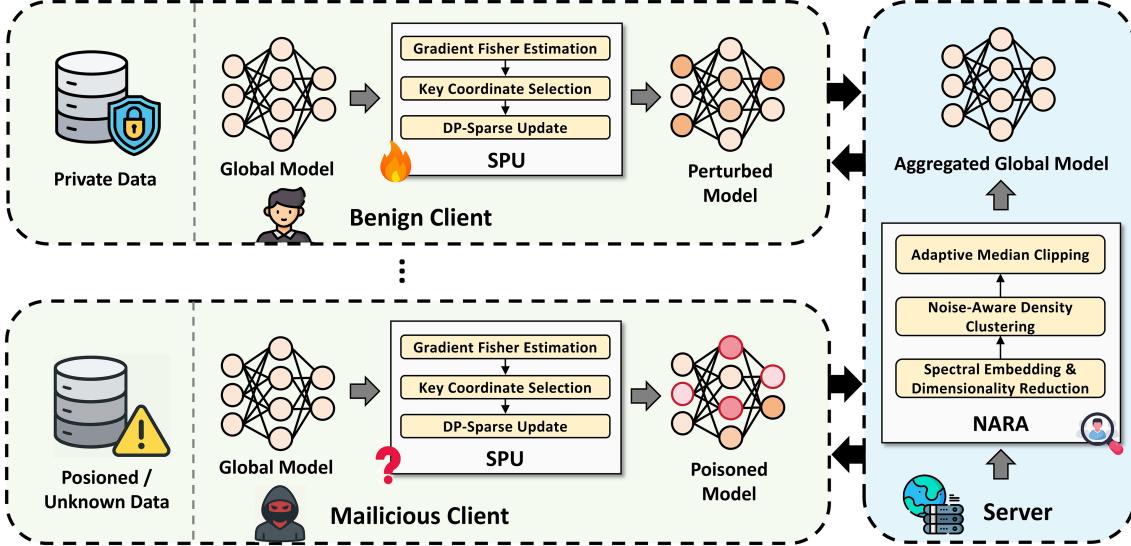


图 1 (网络版彩图) AByzDPFL 的框架图.

Figure 1 (Color online) Framework of AByzDPFL.

算法 1 所示。

**Fisher 信息量估计.** 直觉上, 模型中对预测贡献更大的参数对应更高的信息量, 而 Fisher 信息<sup>[40]</sup>能刻画损失函数曲率、衡量参数对预测结果的敏感度. 故本文采用 Fisher 信息来量化参数可提供的信息量的能力. 一般来说, 观测数据样本  $x$  对模型参数  $\theta$  的 Fisher 信息定义为

$$I(\theta) = \mathbb{E}_{p(x|\theta)} \left[ (\nabla \log p(x|\theta)) (\nabla \log p(x|\theta))^{\top} \right], \quad (1)$$

其中  $\log p(x|\theta)$  为参数  $\theta$  的对数似然函数. 考虑到客户端资源有限, 直接计算该矩阵在深度学习场景下计算与存储开销过大, 故采用经验 Fisher 信息<sup>[41]</sup>近似. 对于客户端  $i$ , 其本地模型  $\mathbf{w}_i$  的经验 Fisher 信息计算为

$$F = \frac{1}{n_i/B} \sum_{i=1}^{n_i/B} \nabla \mathbf{g}(x_i) \cdot \nabla \mathbf{g}(x_i)^{\top}, \quad (2)$$

其中  $n_i$  为本地数据量,  $B$  为批次大小,  $\nabla \mathbf{g}(x_i)$  为样本  $x_i$  的梯度,  $\cdot$  表示各个梯度的外积. 接着, 将每一层梯度的 Fisher 值进行归一化处理, 确保不同尺度参数间的公平比较:

$$\hat{F}_{(l,j)} = \frac{F_{(l,j)} - \min\{F_{(l,j')}\}}{\max\{F_{(l,j')}\} - \min\{F_{(l,j')}\}}, \quad j, j' = 1, \dots, d_l, \quad d = \sum_l d_l, \quad (3)$$

其中  $l = 1, \dots, L$  为层索引,  $j$  为层内参数索引,  $d_l$  为第  $l$  层参数个数. 归一化后的 Fisher 信息  $\hat{F}_{(j)}$  的相对大小反映了各坐标对模型输出及潜在隐私泄露的相对贡献, 用于指导关键参数的选择.

**私有关键坐标选择.** 基于上述信息量估计, 需在保证差分隐私性的前提下有效提取最具代表性的参数坐标, 进一步提高噪声注入效率并兼顾信息保留. 本文在一次性 Laplace 机制 Top- $k$  选择<sup>[42]</sup>的基础上, 创新性地引入动态稀疏率调度. 借鉴余弦退火在学习率调度中的优势<sup>[43]</sup>, 将稀疏率  $p^t$  在训练轮次  $t$  动态调整:

$$p^t = \frac{p}{2} \left( 1 + \cos \left( \frac{\pi \cdot t}{T} \right) \right), \quad (4)$$

其中  $p$  为初始稀疏率,  $T$  为全局通信轮数. 该策略可使模型在训练早期保留更多梯度信息以促进充分探索, 训练后期则聚焦于关键梯度坐标以提升收敛稳定性并减少噪声影响. 随后, 根据当前稀疏率  $p^t$ ,

**算法 1** Selective perturbation update (SPU).

---

**Input:** Current global model  $\mathbf{w}$ , local dataset size  $n$ , local epochs  $E$ , batch size  $B$ , learning rate  $\eta$ , sparsification ratio  $p$ , noise scale  $\sigma$ , clipping norm  $C$ , privacy budget  $(\epsilon_1, \delta_1)$ ;

- 1: Initialize the local model:  $\mathbf{w}_i = \mathbf{w}$ ;
- 2: **for**  $e = 1, \dots, E$  **do**
- 3:   **for** each minibatch  $\mathcal{B}$  **do**
- 4:     **for** each  $x \in \mathcal{B}$  **do**
- 5:       Compute and clip gradient:  $\mathbf{g}(x) \leftarrow \text{Clip}(\nabla f(\tilde{\mathbf{w}}, x), C)$ ;
- 6:     **end for**
- 7:     Aggregate clipped gradients:  $\mathbf{g} = \frac{1}{B} \sum_{x \in \mathcal{B}} \mathbf{g}(x)$ ;
- 8:     // Fisher information estimation
- 9:     Compute empirical Fisher:  $F_{(l,j)} = \frac{1}{n/B} \sum_{x \in \mathcal{B}} \mathbf{g}_j(x) \mathbf{g}_j(x)^\top$ ;
- 10:    Normalize per-coordinate Fisher:  $\hat{F}_{(l,j)} = \frac{F_{(l,j)} - \min\{F_{(l,j')}\}}{\max\{F_{(l,j')}\} - \min\{F_{(l,j')}\}}$ ,  $j, j' = 1, \dots, d_l$ ;
- 11:    // Private key coordinate selection
- 12:    Adjustment of sparsification ratio:  $p^t = \frac{p}{2} (1 + \cos(\frac{\pi \cdot t}{T}))$ ;
- 13:    Determine number of selected coordinates:  $k_l = \lceil p^t \cdot d_l \rceil$ ;
- 14:    Compute Laplace noise scale:  $\lambda = \frac{8 \Delta_F \sqrt{k_l \ln(d_l/\delta_1)}}{\epsilon_1}$ ;
- 15:    Add Laplace noise:  $\tilde{F}_{(l,j)} = \hat{F}_{(l,j)} + \mathbf{u}_{(l,j)}$ ,  $\mathbf{u}_{(l,j)} \sim \text{Lap}(\lambda)$ ;
- 16:    Form key coordinate set:  $\mathcal{K} \leftarrow \sum_l \text{argsort}_{(j)}(\tilde{F}_{(j)})[1 : k_l]$ ;
- 17:    // DP-sparse update
- 18:    **for**  $l = 1, \dots, L$  **do**
- 19:     **for**  $j = 1, \dots, d_l$  **do**
- 20:       **if**  $(l, j) \in \mathcal{K}$  **then**
- 21:         Add Gaussian noise to gradient and update:  $\tilde{\mathbf{w}}_{(l,j)} \leftarrow \mathbf{w}_{(l,j)} - \eta(\mathbf{g}_{(l,j)} + \mathcal{N}(0, \sigma^2))$ ;
- 22:       **else**
- 23:         Retain global parameter:  $\tilde{\mathbf{w}}_{(l,j)} \leftarrow \mathbf{w}_{(l,j)}$ ;
- 24:       **end if**
- 25:     **end for**
- 26:    **end for**
- 27:   **end for**
- 28: **end for**

**Output:** Updated local model  $\tilde{\mathbf{w}}$ .

---

对每层  $l$  计算待选关键坐标数  $k_l$ :

$$k_l = \lceil p^t \cdot d_l \rceil, \quad k = \sum_l k_l. \quad (5)$$

为避免高信息量梯度坐标直接暴露带来的隐私泄露风险, 对层  $l$  中坐标  $j$  的估计 Fisher 值  $\tilde{F}_{(l,j)}$  采样独立 Laplace 噪声进行扰动:

$$\tilde{F}_{(l,j)} = \hat{F}_{(l,j)} + u_{(l,j)}, \quad u_{(l,j)} \sim \text{Lap}(\lambda), \quad (6)$$

其中  $\lambda = \frac{8 \Delta_F \sqrt{k_l \ln(d_l/\delta_1)}}{\epsilon_1}$ ,  $\Delta_F$  为 Fisher 值的敏感度. 然后, 根据该值分层从小到大排序, 选择前  $k_l$  个坐标组成全局关键坐标集合  $\mathcal{K}_l$  ( $\mathcal{K} = \bigcup_l \{(l, j) : j \in \mathcal{K}_l\}$ ), 其满足  $(\epsilon_1, \delta_1)$ -DP. 该过程仅需一次噪声注入与排序, 避免了指数机制<sup>[42,44]</sup>的累计隐私开销  $k\epsilon$ , 保障了后续模型更新的高效性.

**差分隐私稀疏更新.** 本文仅对选中坐标集合  $\mathcal{K}$  进行更新, 以减小有效噪声量的引入, 从而确保模型的可用性. 具体来说, 对高信息量的梯度将注入高斯噪声并更新对应参数, 剩余梯度置为零, 以便在参数更新时直接保持前一次迭代参数. 每个参数坐标  $\mathbf{w}_{(l,j)}$  更新规则如下:

$$\tilde{\mathbf{w}}_{(l,j)} = \begin{cases} \mathbf{w}_{(l,j)} - \eta \tilde{\mathbf{g}}_{(l,j)} = \mathbf{w}_{(l,j)} - \eta (\mathbf{g}_{(l,j)} + \mathbf{b}_{(l,j)}), & (l, j) \in \mathcal{K}, \\ \mathbf{w}_{(l,j)} - \eta \cdot 0 = \mathbf{w}_{(l,j)}, & (l, j) \notin \mathcal{K}, \end{cases} \quad (7)$$

**算法 2** Noise-aware robust aggregation (NARA).

---

**Input:** Number of selected clients  $m$ , local model set  $\{\mathbf{w}_i\}_{i=1}^m$ , noise scale  $\sigma$ , kernel bandwidth  $\gamma$ ;

- 1: // **Spectral embedding and dimensionality reduction**
- 2: Compute similarity matrix:  $R_{ii'} \leftarrow \exp(-\|\mathbf{w}_i - \mathbf{w}_{i'}\|_2^2/(2\gamma^2))$ ,  $i, i' \in \mathcal{A}_t$ ;
- 3: Form degree matrix  $D = \text{diag}(\{D_{ii}\})$ , with  $D_{ii} \leftarrow \sum_{i'} R_{ii'}$ ;
- 4: Compute normalized Laplacian:  $L_{\text{sym}} \leftarrow D^{-1/2}(D - W)D^{-1/2}$ ;
- 5: Eigen-decompose  $L_{\text{sym}} = U\Lambda U^\top$ , take the smallest  $k_0$  eigenvectors to form  $U \in \mathbb{R}^{m \times k_0}$ ;
- 6: // **Noise-adaptive density clustering**
- 7: Set minimum core size:  $\text{MinPts} \leftarrow \lceil m/2 \rceil$ ;
- 8: **for**  $i \in \mathcal{A}_t$  **do**
- 9:   Estimate local density upper bound:  $\text{kDist}_i \leftarrow \text{kDist}(\mathbf{w}_i, \text{MinPts})$ ;
- 10:   Compute noise-tolerant lower bound:  $\text{Eps}_{min} = \sigma\sqrt{2d}$ ;
- 11:   Set adaptive neighborhood radius:  $\text{Eps}_i = \max\{\text{Eps}_{min}, \text{kDist}_i\}$ ;
- 12: **end for**
- 13: Generate candidate benign set:  $\mathcal{Z} \leftarrow \text{DBSCAN}(U, \{\text{Eps}_i\}_{i \in \mathcal{A}_t}, \text{MinPts})$ ;
- 14: // **Adaptive median clipping**
- 15: For each  $i \in \mathcal{Z}$ , compute norm  $r_i = \|\mathbf{w}_i\|_2$ , and let  $r_{\text{med}} = \text{Median}(r_i : i \in \mathcal{Z})$ ;
- 16: **for**  $i \in \mathcal{Z}$  **do**
- 17:   Clip:  $\mathbf{w}_i^c \leftarrow \frac{r_{\text{med}}}{\max\{r_i, r_{\text{med}}\}} \mathbf{w}_i$ ;
- 18: **end for**
- 19: // **Model aggregation**
- 20:  $\mathbf{w}^{t+1} \leftarrow \frac{1}{|\mathcal{Z}|} \sum_{i \in \mathcal{Z}} \mathbf{w}_i^c$ ;

**Output:** Next global model  $\mathbf{w}^{t+1}$ .

---

其中  $\eta$  为学习率,  $\mathbf{b} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$  为均值为 0, 标准差为  $\sigma$  的高斯噪声,  $\mathbf{I}_d$  是  $d$  维单位矩阵. 这种稀疏更新策略显著降低了有效噪声维度, 缓解了全局加噪的维数灾难, 优化了隐私与效用之间的平衡.

## 5.2 噪声感知鲁棒聚合 (NARA)

在差分隐私场景中, 每个本地模型  $\mathbf{w}_i \in \mathbb{R}^d$  都附加了同分布高斯噪声  $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ , 此时模型间的真实几何结构包含两部分信息. (1) 噪声漂移. 所有模型均受到幅度相近的随机扰动. (2) 表征差异. 良性模型间的天然差异与恶意模型的系统性偏移. 服务器端的聚合既要避免因噪声漂移而误判良性模型, 又要能借助表征差异来精准剔除恶意模型, 从而保障联邦学习训练过程的鲁棒性和全局模型的可用性. 为此, 本文提出噪声感知鲁棒聚合 (NARA) 规则, 如算法 2 所示, 包括谱嵌入降维、噪声自适应密度聚类和自适应中值裁剪 3 个步骤.

**谱嵌入降维.** 由于模型参数维度通常较高, 直接在原始参数空间进行模型间距离度量时, 高维随机噪声容易掩盖模型之间真实的语义结构, 使异常检测变得困难. 因此, 首先将客户端上传的本地模型进行谱嵌入降维, 将其映射至低维空间, 以凸显模型之间的整体几何结构, 并有效抑制高维随机噪声对距离计算的干扰. 具体而言, 把每个模型视为图结构中的节点, 节点间权重采用高斯径向基函数定义:

$$R_{ii'} = \exp\left(-\frac{\|\mathbf{w}_i - \mathbf{w}_{i'}\|_2^2}{2\gamma^2}\right), \quad i, i' \in \mathcal{A}_t, \quad (8)$$

其中  $\gamma > 0$  为核带宽, 控制相似度衰减速率. 随后, 构造度矩阵  $D = \text{diag}(D_{ii})$ ,  $D_{ii} = \sum_{i'} R_{ii'}$ , 并据此形成未归一化拉普拉斯矩阵  $L = D - R$ . 进一步, 得到对称归一化拉普拉斯  $L_{\text{sym}} = D^{-1/2}LD^{-1/2}$ , 对其进行特征分解  $L_{\text{sym}} = U\Lambda U^\top$ , 提取最小的  $k_0$  个非平凡特征向量, 组成低维嵌入  $U_{k_0} \in \mathbb{R}^{m \times k_0}$ , 其中  $k_0 \ll d$  为预设的降维嵌入维度. 在降维后的空间中, 随机噪声的扩散性显著降低, 使模型之间真实的几何差异被放大, 从而为后续聚类阶段有效区分良性与恶意模型奠定基础.

**噪声自适应密度聚类.** 在降维后的嵌入空间中, 进一步采用 DBSCAN<sup>[45]</sup> 对模型进行密度聚类. 传统 DBSCAN 以固定邻域半径 Eps 和最小核心近邻数 MinPts 为参数, 只有当某点的 Eps- 邻

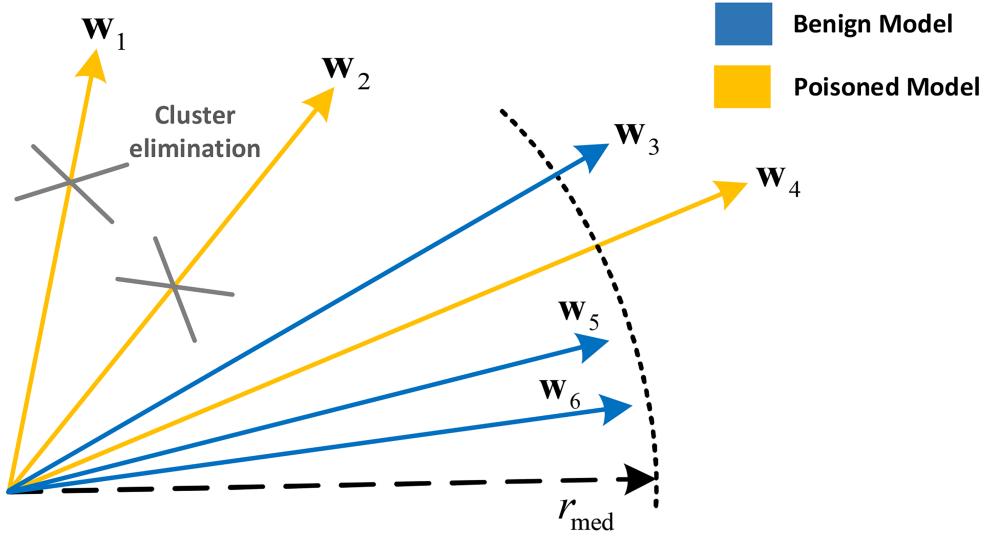


图 2 (网络版彩图) 本地模型集合的几何示意图. 噪声自适应聚类会移除角度偏差较大的模型  $w_1$  和  $w_2$ , 然后中值范数裁剪会抑制剩余角度对齐候选集中高幅值模型  $w_4$  的影响.

**Figure 2** (Color online) Schematic geometric illustration of the local model set. Noise-adaptive clustering removes models with large angular deviations,  $w_1$  and  $w_2$ , and median-norm clipping then suppresses the influence of the high-magnitude model  $w_4$  within the remaining angle-aligned candidates.

域内至少包含 MinPts 个样本时, 该点才被视为核心点, 进而触发连通扩张. 然而, 在差分隐私场景下, Eps 设置过小会把引入噪声的良性模型误判为离群, 过大又会把有毒的偏移模型误并入簇, 导致鲁棒聚合失效. 为此, 本方法结合噪声尺度, 将邻域半径设为每个模型自身的自适应值. 首先对每个模型  $w_i$  计算第 MinPts 个近邻距离  $kDist_i = kDist(w_i, \text{MinPts})$ , 然后根据噪声尺度  $\sigma$  确定聚类半径的下界为两个独立噪声向量的期望距离上界  $Eps_{min}$ , 进而定义每个模型的自适应聚类半径  $Eps_i$  为

$$Eps_{min} = \sigma\sqrt{2d}, \quad Eps_i = \max\{Eps_{min}, kDist_i\}. \quad (9)$$

当模型之间的距离落在  $Eps_{min}$  附近时, 认为其差异主要源于随机噪声, 确保正常噪声波动内的良性模型不会被孤立; 而当某个模型的  $kDist$  显著超出  $Eps_{min}$  时, 则表明其参数偏离了纯噪声范围, 更可能为恶意注入的偏移. 为确保聚类的稳定性与准确性, 本方法还继承了 Mutual-Reachability 距离 [46] 的对称规则, 即当模型间距离满足

$$\|w_i - w_j\| \leq \max\{Eps_i, Eps_j\}, \quad (10)$$

则模型  $w_i$  与  $w_j$  判定为互可达, 从而避免因半径选择不一致带来的聚类偏差. 此外, 考虑到实际场景中恶意客户端数量一般不超过总客户端数量的一半, 本方法将最小核心点数设定为  $\text{MinPts} = \lceil m/2 \rceil$ , 以提取具有稳定几何特征的最大良性模型集合  $\mathcal{Z}$ .

**自适应中值裁剪.** 如图 2 所示, 在聚类后的密集簇内部仍可能存在范数异常的高幅值模型, 即在方向上已对齐但幅值极大. 为抑制此类异常对全局模型的影响, 在良性簇  $\mathcal{Z}$  内引入基于模型范数的中值裁剪策略. 首先, 计算每个候选模型  $w_i$  的  $\ell_2$  范数:  $r_i = \|w_i\|_2 = \sqrt{\sum_{j=1}^d (w_{ij})^2}$ ,  $i \in \mathcal{Z}$ . 取所有范数的中位数  $r_{med} = \text{Median}(\{r_i\}_{i \in \mathcal{Z}})$  作为裁剪阈值, 并对每个模型进行裁剪:

$$w_i^c = \text{Clip}(r_i, r_{med}) = \frac{r_{med}}{\max\{r_i, r_{med}\}} w_i, \quad (11)$$

当  $r_i > r_{med}$  时按比例缩放, 否则保持不变. 最后, 对裁剪后所有  $w_i^c$  取平均作为最终全局模型:

$$w^{t+1} = \frac{1}{|\mathcal{Z}|} \sum_{i \in \mathcal{Z}} w_i^c. \quad (12)$$

**算法 3** Byzantine-robust adaptive differentially private federated learning (AByzDPFL).

**Input:** Number of clients  $N$ , sampling rate  $q$ , number of global rounds  $T$ , number of local epochs  $E$ , selection privacy budget  $(\epsilon_1, \delta_1)$ , update privacy budget  $(\epsilon_2, \delta_2)$ ;

1: Initialize global model  $\mathbf{w}^0$ , sparsification ratio  $p$ ;

2: **for**  $t = 1, \dots, T$  **do**

3:    Sample client subset  $\mathcal{A}^t$  of size  $m$  by probability  $q$ ;

4:    **for** each client  $i \in \mathcal{A}^t$  **in parallel do**

5:     Compute noise scale:  $\sigma = \frac{2C\sqrt{qT \ln(1/\delta_2)}}{|\mathcal{D}_i|^{1/\epsilon_2}}$ ; // Theorem 3

6:     Selective perturbation update:  $\mathbf{w}_i^{t+1} \leftarrow \text{SPU}(i, \mathbf{w}^t, E, p, \sigma, \epsilon_1, \delta_1)$ ; // Algorithm 1

7:    **end for**

8:    Noise-aware robust aggregation:  $\mathbf{w}^{t+1} \leftarrow \text{NARA}(\{\mathbf{w}_i^{t+1}\}_{i \in \mathcal{A}^t}, m, \sigma)$ ; // Algorithm 2

9: **end for**

**Output:** Next global model  $\mathbf{w}^T$ .

### 5.3 整体算法流程

AByzDPFL 的整体流程如算法 3 所示。在  $t$  轮全局通信中，服务器按采样率  $q$  随机选取  $m$  个客户端  $\mathcal{A}^t$ ，并向其分发当前全局模型  $\mathbf{w}^t$ 。每个被选中的客户端根据设定的隐私预算，计算差分隐私噪声尺度  $\sigma$ ，随后执行基于信息量的选择性扰动更新 SPU，生成经过噪声保护的本地模型  $\mathbf{w}_i^{t+1}$ ，并上传至服务器。服务器收集所有客户端模型后，应用噪声感知鲁棒聚合规则 NARA，对上传的本地模型进行异常检测与鲁棒聚合，得到新一轮全局模型  $\mathbf{w}^{t+1}$ 。重复执行该过程，直至达到预设的通信轮数  $T$ 。

## 6 理论分析

### 6.1 隐私性分析

为量化 AByzDPFL 方法在多轮联邦训练过程中的隐私保障水平，本文采用 RDP 框架<sup>[37]</sup>，并结合子采样高斯机制<sup>[47]</sup>和差分隐私的顺序组合性质<sup>[10]</sup>，给出累计隐私损失的严格上界。具体而言，子采样高斯机制被用来分析单次迭代过程中的 RDP 隐私损失上界。

**引理2** (子采样高斯机制<sup>[47]</sup>) 若  $p \leq \frac{1}{5}$ ,  $\sigma \geq 4\Delta$  且  $\alpha$  满足

$$1 \leq \alpha \leq \frac{1}{2}\sigma^2 C_3^2 - 2\ln\sigma, \quad \alpha \leq \frac{\frac{1}{2}\sigma^2 C_3^2 - \ln 5 - 2\ln\sigma}{C_3 + \ln(p\alpha) + 1/(2\sigma^2)}, \quad (13)$$

其中  $C_3 = 1 + \frac{1}{p(\alpha-1)}$ ，则子采样高斯机制对于具有  $\ell_2$  敏感度  $\Delta$  且子采样率为  $p$  的函数， $\mathcal{M}$  满足

$$D_\alpha(\mathcal{M}(\mathcal{D}) \parallel \mathcal{M}(\mathcal{D}')) \leq D_\alpha((1-p)\mathcal{N}_\sigma(0, \Delta^2 \sigma^2 I) + p\mathcal{N}_\sigma(\Delta, \Delta^2 \sigma^2 I) \parallel \mathcal{N}_\sigma(0, \Delta^2 \sigma^2 I)) \leq \frac{2p^2 \Delta^2 \alpha}{\sigma^2}. \quad (14)$$

为了考虑私有坐标筛选时的隐私损失，需结合顺序组合性质。

**性质1** (顺序组合<sup>[10]</sup>) 设随机算法  $\mathcal{M}_i : \mathcal{D} \rightarrow \mathcal{R}_i$  满足  $(\epsilon_i, \delta_i)$ -DP ( $i = 1, 2, \dots, n$ )，则  $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_n) : \mathcal{D} \rightarrow \prod_{i=1}^n \mathcal{R}_i$  满足  $(\prod_{i=1}^n \epsilon_i, \prod_{i=1}^n \delta_i)$ -DP。

进一步分析 AByzDPFL 的整体隐私性如下。

**证明** 首先，对于一对相邻数据集  $\mathcal{D}_i$  和  $\mathcal{D}'_i$ ，单个客户端  $i$  本地更新的敏感度为

$$\Delta_i = \max_{\mathcal{D}_i, \mathcal{D}'_i} \left\| \frac{1}{|\mathcal{D}_i|} \sum_{j=1}^{|\mathcal{D}_i|} \arg \min_{\mathbf{w}_i} f_i(\mathbf{w}_i, \mathcal{D}_{i,j}) - \frac{1}{|\mathcal{D}'_i|} \sum_{j=1}^{|\mathcal{D}'_i|} \arg \min_{\mathbf{w}_i} f_i(\mathbf{w}_i, \mathcal{D}'_{i,j}) \right\| = \frac{2C}{|\mathcal{D}_i|}. \quad (15)$$

接着设算法 1 在第  $t$  次迭代的输出为  $\mathcal{M}(\tilde{\mathbf{w}}_i^t, \mathcal{D}_i)$ ，则在  $\mathcal{D}_i$  和  $\mathcal{D}'_i$  上各自的输出分别为  $\mathcal{M}(\tilde{\mathbf{w}}_i^t, \mathcal{D}_i)$  和  $\mathcal{M}(\tilde{\mathbf{w}}_i^t, \mathcal{D}'_i)$ 。其 Rényi 散度  $D_\alpha(\cdot \parallel \cdot)$  可表示为

$$D_\alpha(\mathcal{M}(\tilde{\mathbf{w}}_i^t, \mathcal{D}_i) \parallel \mathcal{M}(\tilde{\mathbf{w}}_i^t, \mathcal{D}'_i)) = D_\alpha(\mathcal{N}(\tilde{\mathbf{w}}_i^t - \eta \mathbf{g}_i^t, \eta^2 \sigma^2 \mathbf{I}) \parallel \mathcal{N}(\tilde{\mathbf{w}}_i^t - \eta \mathbf{g}'_i^t, \eta^2 \sigma^2 \mathbf{I})). \quad (16)$$

由于两个高斯分布具有相同协方差矩阵, 可以通过平移对比均值差, 并将分布重参数化为中心为零、协方差为  $\sigma^2\mathbf{I}$  的标准形式. 故有

$$D_\alpha(\mathcal{M}(\tilde{\mathbf{w}}_i^t, \mathcal{D}_i) \| \mathcal{M}(\tilde{\mathbf{w}}_i^t, \mathcal{D}'_i)) = D_\alpha(\mathcal{N}(\mathbf{g}_i^t - \mathbf{g}'_i, \sigma^2\mathbf{I}) \| \mathcal{N}(0, \sigma^2\mathbf{I})). \quad (17)$$

可见,  $\tilde{\mathbf{g}}_i^t$  是敏感度为  $\frac{2C}{|D_i|}$  的子采样高斯机制的一个实例. 设子采样率为  $q = \frac{m}{N}$ , 裁剪范数为  $C$ , 噪声标准差为  $\sigma$ . 根据引理 2, 设  $\sigma \geq \frac{8C}{|D_i|}$ ,  $q \leq \frac{1}{5}$ , 可得第  $t$  轮 RDP 损失为

$$\begin{aligned} \epsilon(\alpha)_t &= D_\alpha(\mathcal{M}(\tilde{\mathbf{w}}^t, \mathcal{D}) \| \mathcal{M}(\tilde{\mathbf{w}}^t, \mathcal{D}')) = D_\alpha(\mathcal{N}_\sigma(\bar{\mathbf{g}}^t - \bar{\mathbf{g}}'^t) \| \mu_0) \\ &\leq D_\alpha((1-p)\mu_0 + q\mathcal{N}_\sigma\left(\frac{2C}{|D_i|}\right) \| \mu_0) \leq \frac{8q^2C^2\alpha}{|D_i|^2\sigma^2}, \end{aligned} \quad (18)$$

对  $T$  次迭代求和, 则总 RDP 损失上界为  $\epsilon(\alpha) = \frac{8p^2C^2\alpha T}{|D_i|^2\sigma^2}$ .

由 RDP 到 DP 的转换引理 1 与顺序组合性质 1, 进一步得到以下隐私保证.

**定理3** (隐私性保证) 通过选取  $q \leq \frac{1}{5}$ 、高斯噪声尺度  $\sigma$  满足

$$\sigma \geq \frac{2C\sqrt{qT\log(1/\delta_2)}}{|D_i|\epsilon_2}, \quad (19)$$

则 ABYZDPFL 在  $T$  轮训练后整体满足  $(T\epsilon_1 + \epsilon_2, T\delta_1 + \delta_2)$ -DP.

## 6.2 收敛性分析

为刻画 ABYZDPFL 的全局收敛性质, 令每个客户端  $i \in [N]$  的本地目标为  $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ , 全局目标定义为  $f(\mathbf{w}) = \sum_{i=1}^N p_i f_i(\mathbf{w})$ , 其中  $p_i = \frac{n_i}{\sum_j n_j}$ . 对  $f_i$  作出如下基本假设.

**假设1** ( $L$ - 光滑性)  $f_i$  均为  $L$ - 光滑, 即对于任意  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ , 存在  $L > 0$  使得

$$f_i(\mathbf{x}) \leq f_i(\mathbf{y}) + \nabla f_i(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2. \quad (20)$$

**假设2** (有界梯度)  $f_i$  的梯度在整个参数空间内是有界的. 即存在常数  $G > 0$ , 使得

$$\mathbb{E}\|\nabla f_i(\mathbf{w}_i^t)\| \leq G. \quad (21)$$

**假设3** (随机梯度方差) 每个客户端的随机梯度  $\mathbf{g}_i^t$  对全梯度无偏, 即对于所有  $\mathbf{w}_i \in \mathbb{R}^d$ , 有  $\mathbb{E}[\mathbf{g}] = \nabla f(\mathbf{w}_i)$ , 且其方差有界, 存在常数  $\beta > 0$ , 使得

$$\mathbb{E}\|\mathbf{g}_i^t - \nabla f_i(\mathbf{w}_i^t)\|^2 \leq \beta^2.$$

**证明** 首先给出稀疏选择与高斯差分隐私噪声下的方差界. 在假设 3 和噪声服从均值为 0、标准差为  $\sigma$  的条件下, 客户端的稀疏更新满足

$$\begin{aligned} \mathbb{E}\|\tilde{\mathbf{g}}_i^t - \nabla f_i(\mathbf{w}_i^t)\|^2 &\leq \mathbb{E}\|\tilde{\mathbf{g}}_i^t - \mathbf{g}_i^t\|^2 + \mathbb{E}\|\mathbf{g}_i^t - \nabla f_i(\mathbf{w}_i^t)\|^2 \\ &= \sum_{j=1}^d \mathbb{E}\|\tilde{\mathbf{g}}_{i,j}^t - \mathbf{g}_{i,j}^t\|^2 + \mathbb{E}\|\mathbf{g}_i^t - \nabla f_i(\mathbf{w}_i^t)\|^2 \\ &\leq \beta^2 + k\sigma^2, \end{aligned} \quad (22)$$

其中,  $k$  为经过私有选择的稀疏更新参数个数.

同理, 基于假设 2 推导出稀疏更新后的梯度范数边界为  $\mathbb{E}\|\nabla f_i(\tilde{\mathbf{w}}_i)\|^2 \leq \left(1 - \frac{k}{d}\right)G$ . 记  $f^*$  为全局  $f$  最优值, 则 ABYZDPFL 在非凸条件下的收敛保证如下.

**定理4** (收敛性保证) 在假设 1~3 下, 设每轮本地步数为  $E$ , NARA 筛选后的良性客户端数为  $Z$ , 服务器端中值裁剪阈值为  $C'$ . 定义

$$D = \left(1 - \frac{k}{d}\right)E G + \eta E L G + \frac{C'}{\sqrt{Z}}(\sigma\sqrt{k} + G),$$

$$\Gamma = \frac{E}{Z}(\beta^2 + k\sigma^2) + C'^2 \frac{\sigma^2 k + G^2}{Z}.$$

取步长

$$\eta = \frac{c}{\sqrt{T}}, \quad c \leq \frac{kE}{dL\sqrt{\Gamma + (G + D)^2}},$$

则由 AByzDPFL 在  $T$  轮后的迭代  $\{\mathbf{w}^t\}$  满足

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\mathbf{w}^t)\|^2 \leq \frac{d(f(\mathbf{w}^0) - f^*)}{kE c \sqrt{T}} + GD + \frac{dc}{2kE\sqrt{T}} L(\Gamma + (G + D)^2). \quad (23)$$

## 7 实验评估

### 7.1 实验设置

**实验环境.** 所有实验均部署在 Ubuntu 22.04.2 LTS 系统上, 硬件配置包括 Intel Xeon Gold 5218R @ 2.10 GHz CPU 和 80GB NVIDIA RTX A100 GPU, 基于 PyTorch 1.9.1 框架完成.

**数据集和模型.** 在 MNIST, Fashion MNIST 和 CIFAR-10 三个标准数据集上进行评估. MNIST 数据集包含了 10 类手写数字灰度图像, 共有 70000 张样本, 60000 张用于训练, 10000 张用于测试, 每张图像分辨率为  $28 \times 28$  像素. Fashion MNIST (FMNIST) 数据集同样包含 10 类服装图片, 数据划分与 MNIST 相同, 图像尺寸也为  $28 \times 28$ . CIFAR-10 是一个包含 10 类 RGB 彩色图像的数据集, 每类 6000 张图片, 训练集和测试集分别为 50000 张和 10000 张, 单张图像尺寸为  $32 \times 32$ . 其中, 针对 MNIST 和 FMNIST, 采用经典的卷积神经网络 LeNet-5 进行训练. 对于 CIFAR-10, 则选用 18 层的深度残差网络 ResNet-18.

**基线方法.** 选择与不同类型的基线进行比较, 包括无防御 (no defense): FedAvg [1]; 隐私保护 (privacy): UDP-FL [48] 和 FedLAP-DP [21]; 鲁棒防御 (robustness): Krum [11] 和 FLAME [24]; 隐私和鲁棒兼顾 (privacy & robustness): PRoBit+ [32], DPFL-APA [30], PR-PFL [31] 以及 DPSFL [35].

**部署细节.** 本地训练在独立同分布 (IID) 数据上进行, 使用 Adam 优化器 ( $\beta_1 = 0.9, \beta_2 = 0.999$ ) 和交叉熵损失. 各项超参数设定如下. 全局通信轮数  $T = 100$ , 客户端总数  $N = 100$ , 采样率  $q = 0.6$ , 恶意客户端比例  $\rho = 0.3$ , 批次大小  $B = 50$ , 每轮本地迭代次数  $E = 5$ ; 学习率  $\eta$  在  $\{10^{-2}, 10^{-3}, 10^{-4}\}$  中网格搜索选择; 差分隐私参数  $\epsilon = 10, \delta = 10^{-5}$ , 裁剪阈值  $C = 1.0$ . 对于 AByzDPFL, 设  $\epsilon_1 = 0.1\epsilon/T, \epsilon_2 = 0.9\epsilon, \delta_1 = \delta_2 = 10^{-5}$ , 稀疏比例  $p = 0.6$ , 高斯核带宽  $\gamma = 1.0$  以及降维嵌入维度  $k_0 = 4$ . 基线方法中, Krum 的异常客户端数取  $\rho \cdot m$ , FedLAP-DP 的 MSE 距离项权重为 0.1, DPSFL 部署二级服务器数为 10 并且验证函数采用余弦相似度. 此外, 统一使用主流开源差分隐私库 Opacus 来计算噪声尺度  $\sigma$ .

### 7.2 成员推理攻击下的性能对比

**攻击设置.** 使用成员推理攻击 (membership inference attack, MIA) 来评估各种方法的隐私性, 旨在推断某一条记录是否用于训练目标模型. 基于公开的白盒攻击框架 Whitebox-Attacks<sup>1)</sup> 实现, 攻击者可访问目标模型的内部结构以及部分目标训练数据. 具体而言, 设定成员训练集为原训练集的 20%, 成员测试集为剩余训练数据的 10%, 非成员训练集和测试集均为原测试集的 10%. 每个客户端都独

1) <https://github.com/SPIN-UMass/MembershipWhiteboxAttacks>.

表 2 成员推理攻击模型的结构. 其中,  $\#\text{num\_classes}$  表示类别数,  $\#\text{num\_feds}$  表示参与聚合的模型数量.

**Table 2** Structure of the membership inference attack model. Here,  $\#\text{num\_classes}$  denotes the number of classes, and  $\#\text{num\_feds}$  denotes the number of models involved in aggregation.

Module	Layer type	Description
Gradient	$2 \times$ Fully-Connected + ReLU	$(\#\text{num\_classes} \times 512) \times 1024, 1024 \times 128$
Prediction	$2 \times$ Fully-Connected + ReLU	$\#\text{num\_classes} \times 100, 100 \times 64$
Label	$2 \times$ Fully-Connected + ReLU	$\#\text{num\_classes} \times 128, 128 \times 64$
Correctness	$2 \times$ Fully-Connected + ReLU	$1 \times \#\text{num\_classes}, \#\text{num\_classes} \times 64$
Combine	$3 \times$ Fully-Connected + ReLU	$((128 + 3 \times 64) \times \#\text{num\_feds}), 256 \times 128, 128 \times 64, 64 \times 1$
Output	Sigmoid	—

表 3 不同方法在成员推理攻击下的表现 (%).

Table 3 Performance (%) of different methods in MIA.

Type	Method	MNIST		FMNIST		CIFAR-10	
		Test acc ↑	Attack acc $\xrightarrow{50\%}$	Test acc ↑	Attack acc $\xrightarrow{50\%}$	Test acc ↑	Attack acc $\xrightarrow{50\%}$
No defense	FedAvg [1]	98.52	61.45	89.74	60.76	74.26	60.38
Privacy	UDP-FL [48]	93.85	51.20	83.17	51.33	69.01	51.76
	FedLAP-DP [21]	97.56	51.32	86.25	51.21	71.69	51.65
Robustness	Krum [11]	97.92	60.83	87.65	60.56	72.54	58.85
	FLAME [49]	97.21	60.64	87.71	60.44	72.91	60.14
Privacy & robustness	PRoBit+ [32]	93.29	50.57	83.94	50.61	68.92	50.78
	DPFL-APA [30]	95.91	51.23	86.49	51.24	71.86	51.58
	PR-PFL [31]	95.98	51.57	85.38	51.20	71.14	51.63
	DPSFL [35]	97.63	50.29	87.05	50.45	72.43	50.37
AByzDPFL (Ours)		97.86	51.36	87.12	51.14	72.80	51.50

立、等量地从不重叠的样本中抽取数据, 用于攻击推理模型的训练与评估. 攻击模型的结构如表 2 所示. 输入包括目标模型的最后一层梯度 (gradient)、最后一层预测输出 (prediction)、标签类别 (label) 和正确性 (correctness), 编码层将每个客户端的不同特征拼接再融合, 输出层通过 Sigmoid 函数输出属于训练集的概率.

**结果分析.** 表 3<sup>[1, 11, 21, 30~32, 35, 48, 49]</sup> 给出了 MNIST, FMNIST 和 CIFAR-10 三个数据集上各方法的全局模型测试准确率 (Test acc) 和成员推理攻击成功率 (Attack acc), 其中 Attack acc 越接近 50% 表明隐私保护越好, 而 ↑ 表示测试准确率越高越好. 可以看到, 无防御方法 FedAvg 以及纯鲁棒防御方法 Krum 和 FLAME 的 Attack acc 明显高于随机猜测 50%, 说明它们对 MIA 几乎无防御能力. 纯隐私保护方法 UDP-FL, FedLAP-DP, DPFL-APA 和 PR-PFL 都能将 Attack acc 降至约 51%, 验证了差分隐私噪声注入抵御 MIA 的有效性. 值得注意的是, 在二者兼顾的方法中, DPSFL 借助多级服务器架构、秘密共享协议与二项噪声协同, 取得了最低的 Attack acc, 并在 3 种数据集上分别保持了 97.63%, 87.05% 和 72.43% 的 Test acc. PRoBit+ 通过比特量化策略也获得了与 DPSFL 相近的攻击抵抗效果. 相比之下, AByzDPFL 在选择性扰动更新策略下将 Attack acc 控制在 51.14%~51.50%, 并以 97.86%, 87.12% 和 72.80% 的 Test acc 实现了最低效用损失, 与 FedAvg 相比平均精度仅下降 0.16%~1.18%. 以上结果表明 AByzDPFL 在隐私防护与模型可用性之间达成了最佳平衡.

表 4 不同方法在不同投毒攻击下的全局模型测试准确率 (%).

Table 4 Global model test accuracy (%) of different methods under various poisoning attacks.

Type	Method	MNIST				FMNIST				CIFAR-10			
		LF	SC	MM	MS	LF	SC	MM	MS	LF	SC	MM	MS
No Defense	FedAvg <sup>[1]</sup>	93.02	75.12	67.28	68.53	63.42	37.86	60.50	39.80	59.93	34.62	12.58	24.23
Privacy	UDP-FL <sup>[48]</sup>	87.15	72.30	64.05	62.40	60.33	61.10	54.42	43.60	52.10	38.75	36.50	38.30
	FedLAP-DP <sup>[21]</sup>	90.85	73.30	65.10	68.60	61.55	55.75	58.60	45.70	57.05	40.50	32.40	42.35
Robustness	Krum <sup>[11]</sup>	75.02	90.10	87.46	89.19	65.27	84.37	64.25	35.79	29.56	67.58	13.75	63.56
	FLAME <sup>[24]</sup>	98.66	98.78	98.68	97.79	86.01	86.85	86.74	86.45	73.46	72.71	72.25	72.59
Privacy & Robustness	PRoBit+ <sup>[32]</sup>	91.71	91.01	85.74	85.88	81.97	80.13	74.70	79.81	67.92	65.25	60.80	67.64
	DPFL-APA <sup>[30]</sup>	93.50	93.44	93.10	87.15	83.45	83.61	76.36	75.45	68.64	68.29	63.06	60.75
	PR-PFL <sup>[31]</sup>	93.97	92.66	92.87	91.30	82.34	82.64	78.04	81.18	68.14	67.10	63.45	66.67
	DPSFL <sup>[35]</sup>	96.90	97.04	96.70	96.80	84.65	85.21	84.96	85.05	71.24	71.89	71.35	71.66
AByzDPFL (Ours)		97.65	97.54	97.08	97.25	85.32	86.03	85.52	85.19	72.16	72.00	71.96	72.24

### 7.3 投毒攻击下的性能对比

**攻击设置.**为了评估 AByzDPFL 方法的鲁棒性, 在如下 4 种经典的非定向投毒攻击上进行测试.(1) Label Flipping (LF)<sup>[50]</sup> 是一种经典的数据投毒方式, 攻击者不需要了解训练数据分布就能翻转每个训练样本的标签, 即将原始样本的标签替换为数据集中其他任意类别的标签.(2) Scaling (SC)<sup>[11]</sup> 的原理是扩大恶意模型在全局聚合时所占的比例. 恶意客户端复制部分本地数据并更改为攻击者的目标标签, 接着在增强的数据集上训练本地模型, 之后在上传服务器之前将其乘以缩放因子.(3) Min-Max (MM)<sup>[51]</sup> 希望计算得到的恶意模型与任何其他模型之间的距离最大化, 但又不超过任意两个良性模型更新之间的最大距离.(4) Min-Sum (MS)<sup>[51]</sup> 的目标是使恶意模型与所有良性模型之间的距离平方和, 是任意良性模型与其他良性模型的距离平方和的上限. 在本实验中, 将 Label Flipping 攻击  $d$  标签  $l$  翻转为  $L - l - 1$ , 其中  $L$  是标签总数,  $l \in \{0, 1, \dots, L - 1\}$ , Scaling 攻击设置缩放因子为 8, Min-Max 和 Min-Sum 初始攻击系数均为 30.

**结果分析.**表 4 对比了各方法在 4 种投毒攻击下的全局模型测试精度. 纯鲁棒防御方法 FLAME 虽在所有场景中取得最高精度, 但仅专注于投毒检测, 并未采用任何客户端隐私保护策略. 纯差分隐私方法 UDP-FL 和 FedLAP-DP 在多种投毒场景下表现不佳, 说明单一噪声机制在兼顾隐私和鲁棒性方面的明显不足. 在兼顾隐私与鲁棒性的方案中, AByzDPFL 表现最佳: 在 MNIST 上精度为 97.08%~97.65%, 在 FMNIST 上为 85.19%~86.03%, 在 CIFAR-10 上为 71.96%~72.24%, 与 FLAME 的差距均在 1.5% 以内, 并分别超越 PRoBit+, DPFL-APA 和 PR-PFL. 同时, 在 3 组数据集上的平均精度较 DPSFL 提升约 0.48%, 进一步证明了方法的稳定性与有效性. 总体来看, AByzDPFL 在多数据集和多种投毒攻击下均显著提升了系统鲁棒性, 并在隐私保护与模型性能之间实现了良好平衡.

### 7.4 消融实验

**关键组件测试.**在 CIFAR-10 上, 为评估 AByzDPFL 中关键组件的贡献, 设计 7 种消融变体.(1) w/o SPU, 客户端执行普通无噪声训练.(2) w/o  $\mathcal{K}$ , 不进行关键坐标筛选, 客户端执行标准 DP-SGD.(3) w/o  $p^t$ , 保留关键坐标选择但固定稀疏率, 不使用余弦退火调度.(4) w/o NARA, 服务器端采用标准 FedAvg 而非 NARA.(5) w/o  $Eps_i$ , 将 NARA 中的自适应聚类半径替换为固定半径.(6) w/o  $\mathcal{Z}$ , 在 NARA 中仅执行中值范数裁剪而不做自适应聚类.(7) w/o  $r_{med}$ , 在 NARA 中仅执行自适应聚类而不做中值范数裁剪. 表 5 报告了各配置在成员推理攻击 (MIA,  $\xrightarrow{50}$ ) 成功率及 4 种非定向投毒攻击 (LF/SC/MM/MS,  $\uparrow$ ) 下的全局模型测试准确率, 其中 MIA 越接近 50% 表明隐私保护越好, 而  $\uparrow$  表

表 5 CIFAR-10 上对 AByzDPFL 关键组件的消融实验结果 (%).

Table 5 Ablation results (%) of key components of AByzDPFL on CIFAR-10.

Attack	w/o SPU	w/o $\mathcal{K}$	w/o $p^t$	w/o NARA	w/o $\text{Eps}_i$	w/o $\mathcal{Z}$	w/o $r_{\text{med}}$	AByzDPFL
MIA $\xrightarrow{50}$	60.50	51.25	51.34	51.42	51.53	51.47	51.55	51.50
LF $\uparrow$	72.96	69.18	71.46	38.45	68.94	34.66	68.50	72.16
SC $\uparrow$	72.44	69.10	71.62	12.50	69.67	38.23	45.35	72.00
MM $\uparrow$	72.98	68.04	71.55	15.72	65.10	25.45	65.60	71.96
MS $\uparrow$	72.51	68.53	71.79	18.90	63.32	26.13	68.80	72.24

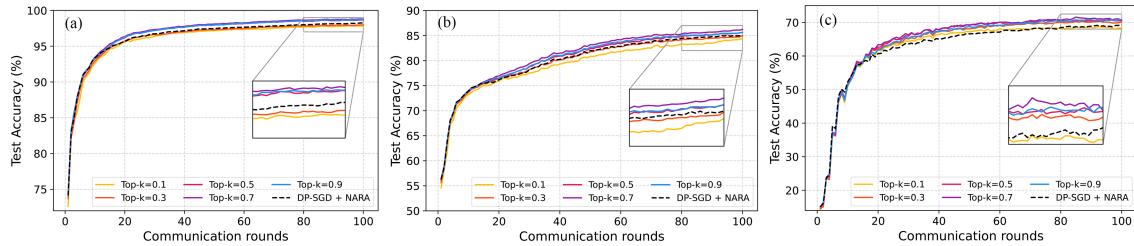


图 3 (网络版彩图) Scaling 攻击下不同稀疏比例的训练曲线. (a) MNIST; (b) FMNIST; (c) CIFAR-10.

Figure 3 (Color online) Training curves of different sparse ratios under Scaling attack. (a) MNIST; (b) FMNIST; (c) CIFAR-10.

示测试准确率越高越好.

从表 5 可见, 取消选择性扰动更新模块 (w/o SPU) 后, MIA 准确率显著上升至 60.50%, 表明本地噪声注入对隐私保护的核心作用. 移除关键坐标筛选 (w/o  $\mathcal{K}$ ) 或固定稀疏率 (w/o  $p^t$ ) 后, MIA 准确率迅速回落至 51.25% 和 51.34%, 而在四种投毒攻击下的测试准确率则下降至 68.04%~69.18% 或 71.46%~71.79%, 说明基于 Fisher 值的 Top- $k$  选择与动态稀疏率调度能够在隐私与性能之间取得平衡. 取消噪声感知鲁棒聚合 (w/o NARA) 时, 在 4 种投毒攻击下的测试准确率骤降至 12.50%~38.45%, 凸显 NARA 对抗有毒更新的必要性; 采用传统固定半径的 DBSCAN 进行聚类 (w/o  $\text{Eps}_i$ ), 表明噪声自适应的半径设计更容易克服噪声的干扰; 仅保留中值裁剪 (w/o  $\mathcal{Z}$ ) 与仅保留聚类 (w/o  $r_{\text{med}}$ ) 均无法兼顾方向性和幅度异常的模型, 前者在 LF/SC 攻击下准确率仅 34.66%~38.23%, 后者对 SC 攻击无力, 进一步证明自适应聚类与范数裁剪两者缺一不可. 完整的 AByzDPFL 在 MIA 和投毒鲁棒性上均取得最优表现, 验证了各子模块及其内部机制的协同增益.

**稀疏比例的影响.** 在 Scaling 攻击下, 分别在 MNIST, FMNIST 和 CIFAR-10 三个数据集上选取不同稀疏比例  $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$  进行实验. 由图 3 可见, 当稀疏比例较低 (如  $p = 0.1$ ) 时, 模型训练初期进展缓慢, 即使迭代至 100 轮, 最终准确率也明显落后于其他设置. 将稀疏比例提高到 0.5 能够一定程度上兼顾收敛速度和精度, 但依然逊于更高稀疏率. 值得关注的是, 当  $p = 0.7$  时, 模型不仅在早期阶段表现出较快的收敛速度, 而且在收敛末期能够有效抑制噪声带来的影响, 最终在 MNIST, FMNIST 和 CIFAR-10 上分别达到 97.77%, 86.22% 和 71.52% 的最优准确率, 显著优于全量更新 (DP-SGD+NARA) 和其他稀疏比例的结果. 尽管  $p = 0.9$  的曲线在收敛速度上与  $p = 0.7$  相近, 但由于过高的参数更新比例导致噪声累积, 其最终性能略有下降. 尤其是在 CIFAR-10 等更具挑战性的任务上, 适度稀疏更新 ( $p = 0.7$ ) 所带来的优势更加突出. 整体来看, 稀疏比例的合理设定不仅有助于抑制冗余噪声扩散, 还能最大限度地保留关键梯度信息, 从而在隐私保护和鲁棒性兼容的高要求场景下, 取得更优的泛化性能.

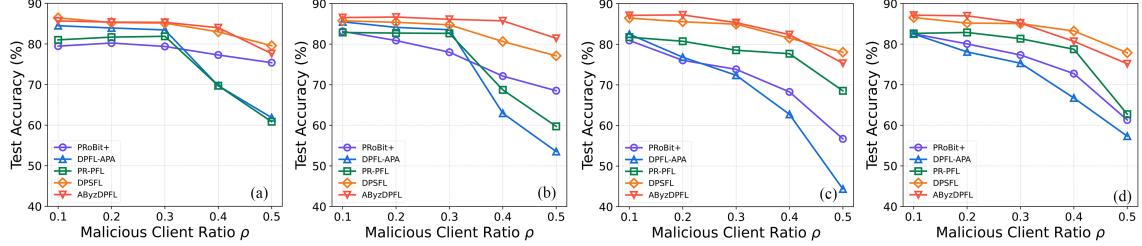


图 4 (网络版彩图) 不同恶意客户端比例下的全局测试准确率. (a) Label Flipping attack; (b) Scaling attack; (c) Min-Max attack; (d) Min-Sum attack.

**Figure 4** (Color online) Global test accuracy under different malicious client ratios. (a) Label Flipping attack; (b) Scaling attack; (c) Min-Max attack; (d) Min-Sum attack.

## 7.5 性能与开销对比

**不同恶意客户端比例.** 在 FMNIST 上进一步评估了 AByzDPFL 方法在不同恶意客户端比例下的性能. 其中, 恶意客户端的比例  $\rho$  从 0.1 递增到 0.5, 中间间隔为 0.1, 同样实施 4 种投毒攻击, 其他实验设置采用默认值. 实验结果如图 4 所示. 整体来看, AByzDPFL 在最轻度攻击 ( $\rho = 0.1$ ) 下即可保持超过 85% 的测试精度, 并随着恶意比例增大出现逐步下降. 当  $\rho = 0.4$  和  $\rho = 0.5$  时, Label Flipping 和 Scaling 攻击下的精度分别降至约 84.00% 和 85.74%, Min-Max 与 Min-Sum 下的最低精度分别为 82.38% 和 80.68%. 在高攻击比例条件下, DPSFL 在多个场景中稍微超越 AByzDPFL, 例如 Label Flipping 情况下  $\rho = 0.5$  时 DPSFL 为 79.57%, 而 AByzDPFL 为 77.61%, Min-Sum 也以 77.87% 略高于 75.17%. 不过, 两者在  $\rho \geq 0.4$  时均出现较大回落, 跌幅接近 10%. 相比之下, PR-PFL, DPFL-APA 和 PRoBit+ 在  $\rho = 0.4$  及以上场景中精度均急剧下滑, 大多跌破 70%, 难以满足高安全性的实际部署需求. 由此可见, 当恶意客户端比例较高时, AByzDPFL 与 DPSFL 能较好地兼顾隐私保护与鲁棒聚合, 而 AByzDPFL 在中低比例阶段展现出更小的性能退化, 体现了其在不同恶意比例下的稳定性和实用性.

**不同 Non-IID 程度.** 在 FMINST 上探究数据非独立同分布程度对不同防御方法的影响. 通过调节狄利克雷 (Dirichlet) 分布参数  $\alpha$  (从 0.1 到 1.0) 来控制分布的浓度或客户端数据的异质性,  $\alpha$  越小, 数据分布越极端, 每个客户端类别分布越单一, Non-IID 程度越高, 反之亦然. 实验结果如图 5 所示. 可以看到, AByzDPFL 在最极端的 Non-IID 情况 ( $\alpha = 0.1$ ) 下依然能够保持 78.03%~79.24% 的测试精度, 而 DPSFL 在相同条件下仅为 77.68%~80.33%. 随着  $\alpha$  增大, 各方法精度普遍上升, 并在近 IID 条件 ( $\alpha = 0.9$ ) 下收敛至 84.42%~86.11%. 与此形成对比的是, PRoBit+ 和 DPFL-APA 在  $\alpha \leq 0.3$  时出现了超过 5% 的明显抖动, PR-PFL 也表现出不稳定的波动. 总体来看, 当客户端的本地数据分布差异较大时, AByzDPFL 可以弱化该异质性困难, 有效检测并减轻投毒攻击的影响, 聚合出较为健壮的全局模型.

**开销对比.** 表 6 列出了各方法在 FMNIST 上单轮全局更新的计算与通信开销. 在默认实验设置中, 每轮选定 60 个客户端, 每个客户端 ( $C_i$ ) 上传 0.17 MB 的 LeNet5 模型, 服务器 ( $S$ ) 下发广播量为  $0.17 \times 60 \approx 10.16$  MB. FedAvg [1] 最轻量, 客户端耗时 0.52 s, 服务器广播仅 0.01 s, 但缺乏隐私与鲁棒机制. PRoBit+ [32] 使用一比特量化使上传延迟增至 0.69 s, 解量化为 0.14 s, 通信量保持不变, 但前文实验结果显示该方法的精度损失较大. DPSFL [35] 通过 10 个二级服务器 (2ndSvrs) 执行秘密共享和噪声注入, 带来了 14.23 s 的额外延迟. DPFL-APA [30] 在边缘节点进行异常检测, 增加了 3.27 s 的延迟. PR-PFL [31] 在客户端执行本地 DP-SGD (0.76 s) 并在服务器基于聚类进行鲁棒聚合 (1.45 s). AByzDPFL 通过经验 Fisher 值计算实现稀疏更新, 客户端计算耗时 0.91 s, 服务器端通过仅在最后一层进行降维和聚类优化, 聚合耗时 1.19 s. 总体而言, AByzDPFL 实现了隐私保护与鲁棒聚合的最佳折中, 保持了可控的延迟和通信开销.

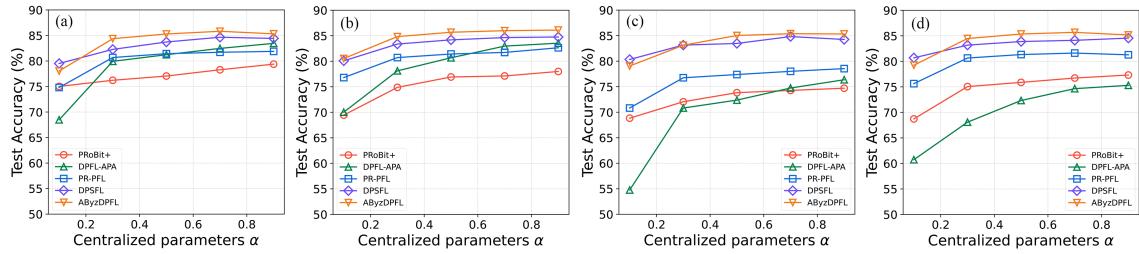


图 5 (网络版彩图) 不同 Non-IID 下的全局模型测试准确率. (a) Label Flipping attack; (b) Scaling attack; (c) Min-Max attack; (d) Min-Sum attack.

Figure 5 (Color online) Global model test accuracy under different Non-IID conditions. (a) Label Flipping attack; (b) Scaling attack; (c) Min-Max attack; (d) Min-Sum attack.

表 6 FMNIST 上每次全局轮数的通信与时间开销对比.

Table 6 Comparison of communication and time overhead of per global round on FMNIST.

Method	FedAvg [1]		PRoBit+ [32]		DPSFL [35]		DPFL-APA [30]		PR-PFL [31]		AByzDPFL	
	$C_i$	$\mathcal{S}$	$C_i$	$\mathcal{S}$	$C_i$	2ndSvrs	$C_i$	EN	$C_i$	$\mathcal{S}$	$C_i$	$\mathcal{S}$
Time (s)	0.52	0.01	0.71	0.14	0.70	14.23	1.13	0.70	3.27	0.41	0.76	1.45
Params (MB)	0.03	10.16	0.17	10.16	1.69	0.17	10.16	0.17	10.16	0.17	10.16	0.17

## 8 总结

针对联邦学习差分隐私保护和拜占庭鲁棒聚合之间的固有矛盾,本文提出了一种自适应拜占庭鲁棒差分隐私联邦学习方法 AByzDPFL. 该方法通过基于 Fisher 信息的关键坐标选择机制,在本地差分隐私下实现稀疏更新,从而显著降低噪声注入的维度,减轻其对异常检测和全局聚合性能的不利影响. 在聚合端,通过降维与噪声自适应聚类协同,有效剔除偏离良性分布的有毒更新,并结合中值范数裁剪抑制高幅值残余异常,从而提升了高维噪声扰动场景下的聚合鲁棒性. 理论分析和系列对比实验验证了所提方法的有效性,并表明其在恶意客户端比例和高度 Non-IID 条件下仍具良好适应性. 未来工作可扩展至大规模 Non-IID、设备异构及通信受限等实际应用场景,并考虑将所提方法与前沿的联邦优化算法、可信执行环境、区块链等技术深度融合,从而进一步提升联邦学习系统的安全性和实用性. 此外,还需关注完全不可信环境中潜在的恶意服务器问题,重点研究聚合结果的可验证机制,完善端到端的安全保障体系.

致谢 本论文研究成果由未来区块链与隐私计算高精尖创新中心建设资助.

## 参考文献

- McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of Artificial Intelligence and Statistics, 2017. 1273–1282
- Gao S, Yuan L P, Zhu J M, et al. A blockchain-based privacy-preserving asynchronous federated learning. Sci Sin Inform, 2021, 51: 1755–1774 [高胜, 袁丽萍, 朱建明, 等. 一种基于区块链的隐私保护异步联邦学习. 中国科学: 信息科学, 2021, 51: 1755–1774]
- Zhang Q N, Zhu J M, Gao S, et al. Incentive mechanism for federated learning based on blockchain and Bayesian game. Sci Sin Inform, 2022, 52: 971–991 [张沁楠, 朱建明, 高胜, 等. 基于区块链和贝叶斯博弈的联邦学习激励机制. 中国科学: 信息科学, 2022, 52: 971–991]
- Yu J X, Shi R H. DDoS attack detection in the Internet of Vehicles based on reinforced federated learning. Sci Sin Inform, 2025, 55: 1221–1238 [于俊骁, 石润华. 基于强化联邦学习的车联网 DDoS 攻击检测. 中国科学: 信息科学, 2025, 55: 1221–1238]

- 5 Zhang H D, Yang L, Yu J, et al. Federated continual learning based on prototype learning. *Sci Sin Inform*, 2024, 54: 2428–2442 [张浩东, 杨柳, 于剑, 等. 基于原型学习的联邦持续学习方法. 中国科学: 信息科学, 2024, 54: 2428–2442]
- 6 Zarifzadeh S, Liu P, Shokri R. Low-cost high-power membership inference attacks. In: Proceedings of the 41st International Conference on Machine Learning, 2024. 235: 58244–58282
- 7 Liu H, Wu Y, Yu Z, et al. Please tell me more: privacy impact of explainability through the lens of membership inference attack. In: Proceedings of IEEE Symposium on Security and Privacy (SP), 2024. 4791–4809
- 8 Xu Z, Jiang F, Niu L, et al. ACE: a model poisoning attack on contribution evaluation methods in federated learning. In: Proceedings of the 33rd USENIX Security Symposium (USENIX Security 24), 2024. 4175–4192
- 9 Zhuang H, Yu M, Wang H, et al. Backdoor federated learning by poisoning backdoor-critical layers. In: Proceedings of the 2024 International Conference on Learning Representations (ICLR), 2024
- 10 Dwork C. Differential privacy. In: Proceedings of International Colloquium on Automata, Languages, and Programming, 2006. 1–12
- 11 Blanchard P, Mhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In: Proceedings of Advances in Neural Information Processing Systems, 2017
- 12 Yao A C. Protocols for secure computations. In: Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (SFCS 1982), 1982. 160–164
- 13 Rivest R L, Adleman L, Dertouzos M L, et al. On data banks and privacy homomorphisms. *Found Secure Comput*, 1978, 4: 169–180
- 14 Geyer R C, Klein T, Nabi M. Differentially private federated learning: a client level perspective. 2017. ArXiv:1712.07557
- 15 McMahan H B, Ramage D, Talwar K, et al. Learning differentially private recurrent language models. 2017. ArXiv:1710.06963
- 16 Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016. 308–316
- 17 Wei K, Li J, Ding M, et al. Federated learning with differential privacy: algorithms and performance analysis. *IEEE Trans Inform Forensic Secur*, 2020, 15: 3454–3469
- 18 Andrew G, Thakkar O, McMahan B, et al. Differentially private learning with adaptive clipping. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34: 17455–17466
- 19 Cheng A, Wang P, Zhang X S, et al. Differentially private federated learning with local regularization and sparsification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 10122–10131
- 20 Xu Z, Collins M, Wang Y, et al. Learning to generate image embeddings with user-level differential privacy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 7969–7980
- 21 Wang H P, Chen D, Kerkouche R, et al. FedLAP-DP: federated learning by sharing differentially private loss approximations. PoPETs, 2024, 2024: 372–390
- 22 Wang B, Dai X R, Wang W, et al. Adversarial examples for poisoning attacks against federated learning. *Sci Sin Inform*, 2023, 53: 470–484 [王波, 代晓蕊, 王伟, 等. 面向联邦学习的对抗样本投毒攻击. 中国科学: 信息科学, 2023, 53: 470–484]
- 23 Yin D, Chen Y, Kannan R, et al. Byzantine-robust distributed learning: towards optimal statistical rates. In: Proceedings of International Conference on Machine Learning, 2018. 5650–5659
- 24 Fung C, Yoon C J, Beschastnikh I. The limitations of federated learning in sybil settings. In: Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020), 2020. 301–316
- 25 Li X, Qu Z, Zhao S, et al. LoMar: a local defense against poisoning attack on federated learning. *IEEE Trans Dependable Secure Comput*, 2021, 20: 437–450
- 26 Nguyen T D, Rieger P, de Viti R, et al. FLAME: taming backdoors in federated learning. In: Proceedings of the 31st USENIX Security Symposium (USENIX Security 22), 2022. 1415–1432
- 27 Bao W, Wu J, He J. BOBA: Byzantine-robust federated learning with label skewness. In: Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS), 2024
- 28 Cao X, Fang M, Liu J, et al. FLTrust: Byzantine-robust federated learning via trust bootstrapping. 2020. ArXiv:2012.13995
- 29 Chu T, Garcia-Recuero A, Iordanou C, et al. Securing federated sensitive topic classification against poisoning attacks. In: Proceedings of Network and Distributed System Security Symposium (NDSS), 2023
- 30 Zhou J, Wu N, Wang Y, et al. A differentially private federated learning model against poisoning attacks in edge computing. *IEEE Trans Dependable Secure Comput*, 2022, 20: 1941–1958

- 31 Yang R, Shen X, Xu C, et al. PR-PFL: a privacy-preserving and robust personalized federated learning framework. In: Proceedings of International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2024. 163–170
- 32 Lan M, Xiao S, Zhang W. One-bit model aggregation for differentially private and byzantine-robust personalized federated learning. 2025. ArXiv:2507.03973
- 33 Tang X, Shen M, Li Q, et al. PILE: robust privacy-preserving federated learning via verifiable perturbations. IEEE Trans Dependable Secure Comput, 2023, 20: 5005–5023
- 34 Rathee M, Shen C, Wagh S, et al. ELSA: secure aggregation for federated learning with malicious actors. In: Proceedings of IEEE Symposium on Security and Privacy (SP), 2023. 1961–1979
- 35 Zhang C, Weng J, Weng J, et al. Robust and secure federated learning with verifiable differential privacy. IEEE Trans Dependable Secure Comput, 2025, 22: 5713–5729
- 36 Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd Theory of Cryptography Conference, 2006. 265–284
- 37 Mironov I. Rényi differential privacy. In: Proceedings of IEEE 30th Computer Security Foundations Symposium (CSF), 2017. 263–275
- 38 Zhou Y, Wu Z S, Banerjee A. Bypassing the ambient dimension: private SGD with gradient subspace identification. In: Proceedings of the 2021 International Conference on Learning Representations (ICLR), 2021
- 39 Wang C, Zhu Y, Su W J, et al. Neural collapse meets differential privacy: curious behaviors of NoisyGD with near-perfect representation learning. In: Proceedings of International Conference on Machine Learning, 2024
- 40 Amari S I. Natural gradient works efficiently in learning. Neural Comput, 1998, 10: 251–276
- 41 Martens J. Deep learning via Hessian-free optimization. In: Proceedings of International Conference on Machine Learning, 2010. 27: 735–742
- 42 Qiao G, Su W, Zhang L. Oneshot differentially private top-k selection. In: Proceedings of International Conference on Machine Learning, 2021. 8672–8681
- 43 Liu S, Yin L, Mocanu D C, et al. Do we actually need dense over-parameterization? In-time over-parameterization in sparse training. In: Proceedings of International Conference on Machine Learning, 2021. 6989–7000
- 44 Durfee D, Rogers R M. Practical differentially private top-k selection with pay-what-you-get composition. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 32
- 45 Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of KDD, 1996. 96: 226–231
- 46 Campello R J, Moulavi D, Zimek A, et al. Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Trans Knowl Discov Data, 2015, 10: 1–51
- 47 Mironov I, Talwar K, Zhang L. Rényi differential privacy of the sampled Gaussian mechanism. 2019. ArXiv:1908.10530
- 48 Wei K, Li J, Ding M, et al. User-level privacy-preserving federated learning: analysis and performance optimization. IEEE Trans Mobile Comput, 2021, 21: 3388–3401
- 49 Zhang H, Li X, Xu M, et al. BADFL: backdoor attack defense in federated learning from local model perspective. IEEE Trans Knowl Data Eng, 2024, 36: 5661–5674
- 50 Fang M, Cao X, Jia J, et al. Local model poisoning attacks to byzantine-robust federated learning. In: Proceedings of the 29th USENIX Security Symposium (USENIX Security 20), 2020. 1605–1622
- 51 Shejwalkar V, Houmansadr A. Manipulating the byzantine: optimizing model poisoning attacks and defenses for federated learning. In: Proceedings of NDSS, 2021

# Adaptive Byzantine-robust differentially private federated learning

Yuhua WANG<sup>1</sup>, Qinnan ZHANG<sup>1\*</sup>, Wangjie QIU<sup>1</sup>, Zichuan CHAI<sup>2</sup>, Sheng GAO<sup>3</sup>, Jianming ZHU<sup>3</sup>, Yongxin TONG<sup>4</sup> & Zhiming ZHENG<sup>1</sup>

1. Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, Institute of Artificial Intelligence, Beihang University, Beijing 100191, China

2. School of Mechanical Engineering & Automation, Beihang University, Beijing 100191, China

3. School of Information, Central University of Finance and Economics, Beijing 100081, China

4. School of Computer Science and Engineering, Beihang University, Beijing 100191, China

\* Corresponding author. E-mail: zhangqn@buaa.edu.cn

**Abstract** Federated learning (FL) enables collaborative training across devices while keeping data local. In practice, however, it faces two security bottlenecks: privacy leakage and poisoning attacks. While differential privacy (DP) and Byzantine-robust aggregation are effective in their respective domains, their coupling entails an inherent conflict: DP noise inflates the variance of benign updates and simultaneously masks the systematic shifts of malicious ones, making them hard to distinguish. To address this, we propose adaptive Byzantine-robust differentially private federated learning (AByzDPFL), which aims to improve distinguishability by reducing the noise dimension and amplifying the geometric differences between models. On the client side, we adopt a Fisher-information-based private selection mechanism that dynamically chooses key parameter coordinates. Noise is injected only within this low-dimensional subspace, which reduces the effective noise dimension and lowers the variance of benign models. On the server side, spectral embedding highlights the intrinsic geometric structure, followed by a noise-scale-adaptive clustering radius that includes noise-perturbed benign models while filtering systemic shifts beyond the noise range. Additionally, we apply adaptive median-norm clipping to suppress high-magnitude anomalous updates within the cluster. We establish upper bounds on privacy loss and convergence, and experiments show that AByzDPFL strikes a balance between privacy and robustness while outperforming existing mainstream baselines.

**Keywords** federated learning, differential privacy, Byzantine robustness, selective update, noise adaptation